预印本与期刊相似主题热点对比分析
——以新冠疫情主题为例

杜潇霖[1]，虞为[1*]，陈俊鹏[2]

（1.江苏省数据工程与知识服务重点实验室，南京大学信息管理学院，南京，210023）

（2.南京财经大学信息工程学院，南京，210023）

摘　要：【目的】对比分析预印本与期刊相似主题的区别与联系，以新冠疫情主题为例证明二者在研究热点与侧重上存在互补。【方法】本文采用 BERTopic 主题分析模型，并提出了二维象限主题对比法，从热度、影响两个维度对预印本与期刊两个来源的相似热点主题进行对比。【结果】分析 bioRxiv、medRxiv、Scopus 平台上新冠疫情相关文献共约七万条，对比预印本和期刊各个维度的主题分布情况，发现以下规律：1）主题热度方面，预印本更偏向医学层面，期刊更偏向社会层面；2）主题影响方面，预印本更偏向预防主题，期刊更偏向病毒传播主题；3）预印本和期刊在热度和影响力方面有部分重回主题，如疫苗相关主题；4）预印本中有期刊中所没有的特有的热点主题，如病毒变异等。【结论】总体来看，预印本的主题更偏向理论与学术，而期刊主题更偏向实践与社会；预印本主题在原理方面分支更细致，而期刊主题在社会层面涉及面更广，二者在内容方面可以形成有效互补。

关键词：预印本；Scopus；主题分析；热点主题；BERTopic；COVID-19

Comparative analysis of hot topics between preprints and periodicals

——A case study of COVID-19

Du Xiaolin, Yu Wei, Chen Junpeng

(1. Jiangsu Key Laboratory of data engineering and knowledge service, School of information management, Nanjing University, Nanjing, 210008, China) (2. School of information engineering, Nanjing University of Finance and Economics, Nanjing, 210023, China)

Abstract: [Objective] To compare and analyze the differences and connections between preprints and periodicals on similar topics, and take the COVID-19 topic as an example to prove that the two are complementary in research hotspot and emphasis. **[Methods]** In this paper, the BERTopic topic

*本文系国家社科基金项目“协同共治视角下图书馆推进开放科学的服务模式研究”（项目编号：21BTQ030）研究成果之一

*虞为是本篇论文的通讯作者（E-mail：yuwei@nju.edu.cn）。

analysis model was adopted and the two-dimensional quadrant method was combined to compare the similar hot topics of preprint and periodical sources from two dimensions of heat and influence. **[Results]** About 70,000 literatures related to COVID-19 on bioRxiv, medRxiv and Scopus were analyzed. The following rules were found: 1) In terms of topic popularity, preprints were more medical, while periodicals were more social; 2) In terms of topic impact, preprints are more likely to focus on prevention, while periodicals are more likely to focus on virus transmission; 3) Preprints and periodicals have partially duplicated topics in terms of popularity and impact, such as vaccine-related topics; 4) There are special hot topics in the preprint that are not in the periodical, such as virus mutation. **[Conclusion]** In general, the preprint topics are more theoretical and academic, while the periodical topics are more practical and social. The preprint topic is more detailed in principle, while the periodical topic is more extensive in society, and the two can effectively complement each other in content.

Keywords: Preprint; Scopus; Topic analysis; Hot topic; BERTopic; COVID-19

1 引言

随着开放科学的发展，预印本（Preprints）作为一种区别于学术期刊的科学出版物逐渐被更多读者认可。预印本是指未在同行评议的学术期刊上正式出版的科研论文手稿[1]。根据 Ross 等人的研究，自 1992 年开放科学运动的兴起，预印本的发展进入扩张期[2]，以学界自治为基础的预印本学术交流模式正逐渐改变传统的以期刊为主体的单一学术交流模式[3]。唐耕砚提出预印本平台缩短了科学交流时滞，促进科学交流体系的去中心化[4]，认为预印本改变了学术出版模式，繁荣了学术交流[5]。预印本与期刊相辅相成、优势互补，共同构成当下重要的学术交流途径，许多著名的研究成果，如庞加莱猜想的证明[6]、Google 的 BERT 模型[7]等，甚至专门发表在预印本平台。

大量预印本及其特点相关的研究归纳了预印本在学术交流中的作用与优势所在。刘菊红[8]通过案例分析，认为开放获取、先见优势、质量歧视等特点使得有预印本的论文更具引用优势。徐诺[9]等人总结了预印本的特点，即时效性强、开放获取、评审透明多元等，认为预印本可以发掘优质稿源、缩短评审周期、创新评审方式。唐耕砚[4]认为预印本可以弥合科学交流的时滞鸿沟，也是强调了速

度优势。周阳[10]总结了预印本的四个优势，即避免审稿偏见、发表速度快、可修改、提供首发证明。汪庆[11]等分析多个主流平台的预印本，认为预印本有出版速度快、开放融合、格式灵活、审核指标多维等特点。总而言之，现有研究提及预印本的特点与优势，主要有两大类：一是没有同行评议带来的速度优势，即弥合时滞，方便交流；二是发表形式灵活带来的内容优势，即开放获取，公开评议，避免偏见，这些优势使得预印本在学术交流中发挥重要作用。

然而，现有研究大多强调预印本的速度优势能够与期刊的质量优势形成互补，却忽视对预印本内容的深入研究。预印本没有同行评议的特点不仅带来发表速度快的优势，也使得内容更加自由，更容易出现创新的观点，但质量隐患一直是阻碍预印本内容相关研究的绊脚石，主要原因是非学术或伪学术论文混入预印本平台的现象，影响了公众对预印本的信任度[12]。随着预印本平台质量控制机制的完善[10, 12, 13]，预印本的质量已经今非昔比，许多研究也开始研究预印本的内容特征，并论证其可靠性。主要方法是对比预印本与其正式出版版本的修改情况，比较一致性、关键研究特征、可解释性等[14-16]。然而，这些研究旨在证明预印本与期刊的联系，却忽视了预印本与期刊整体的区别。目前仍然缺乏对预印本和期刊在相似主题中研究内容、研究热点、影响力等方面的深入研究。

要对比预印本与期刊，文献数量较多、能开放获取的新型冠状病毒肺炎疫情（COVID-19）相关主题是一个很好的研究案例。在新冠疫情期间，以 bioRxiv、medRxiv 为代表的医学预印本平台适应了对学术成果快速、便捷交流的需求，迎来发展的高峰期。刘春丽[17]等人证明了 bioRxiv 自存档在被引次数、社会关注度、临床转化潜力方面的优势，更多科学家选择率先将成果发表在预印本平台上，因此疫情主题的预印本数量较多。期刊方面，医学领域的出版机构与平台出台大量政策，鼓励学术成果的开放获取，缩短出版流程。Homolak 在研究中统计 PubMed、Scopus 平台上 2020 年的文章发表时滞，发现 COVID-19 相关文献发表时滞极短[18]，因此疫情主题的期刊文献数量较多且便于获取。要对比研究预印本与期刊，选择能够开放获取且数量较多的新冠疫情主题的文献作为研究对象是合理的。

疫情期间的两个主要的学术交流途径：预印本与期刊，在发表速度与研究质量方面各有所长，但关于其内容方面差异的研究较少。探究预印本与期刊在相似研究主题上的联系与差异，归结各自的偏好与特长所在，可以更好地发挥二者在

学术交流中的作用。对此，本文以新冠疫情主题为例，选取 bioRxiv、medRxiv、Scopus 平台文献作为预印本与期刊的代表，使用 BERTopic 主题分析方法，从热度、影响两个维度，对比预印本和期刊各自偏向的研究主题，探索预印本的特点所在，为期刊预印本更好地合作互补提供参考。

2 相关工作

目前关于 COVID-19 文献的研究，主要将研究重点放在传统文献计量上。2020 年，匡登辉[19]等人以 COVID-19 预印本为研究对象，主要从时间分布、学术与社会影响力等方面进行分析，发现其发文量在 2020 年 1 月 26 日后开始增长、高影响力预印本中 bioRxiv 平台的文献占比较大等规律。同年李爱花[20]等人基于文献计量学方法分析 COVID-19 研究现状，针对国内外期刊、预印本多种数据源，进行时间分布、国家机构分布、研究热点关键词聚类等分析。该研究在传统文献计量学方面较为全面，但对内容方面的分析只有关键词聚类，且并未研究预印本文献。2022 年，Santos[21]等人基于科学计量指标，从研究人员、机构等角度，对 COVID-19 科研成果进行描述上与时间上的分析，发现期刊仍是最常见来源，而预印本的使用比例也越来越多。现有研究在传统文献计量学方面的分析较为全面，无论是时间分布、期刊分布、国家分布还是影响力，都可以为研究人员提供参考，但对文献的文本内容与主题的研究较少，更多的是以推特等社交媒体文本为研究对象，方法也多是情感分析，如接种疫苗的公众观点分析[22]等。对于疫情相关文献，尤其是预印本的内容与主题仍需进一步研究。

关于某一领域的热点主题的研究已经较为成熟，目前主要有两种研究思路，一是传统文献计量学方法，二是主题模型。

传统文献计量学方法主要有词频分析、共词分析、热词分析等。词频分析简单地考虑词频，忽略了词的含义及联系，也就无法判断哪些词归属于一个主题；共词分析则考虑了文献集中词汇共现的情况，通过构建共词矩阵，进一步通过聚类等方法识别主题[23]。热词发现的方法大致分为两类，一是基于规则的方法，即结合领域专业知识构建规则来进行识别，如王志涛等基于词典与规则分析微博文本的方法[24]；二是基于统计的方法，需要在语料库上进行训练，较前一种方法可移植性强，但存在准确率方面的不足[25]。但这些方法都存在两个共有的缺点，一是片面强调高频词，二是忽略了文本的语义信息。

主题模型可以很好地弥补传统文献计量学方法的不足，也可以分为两大类：一是基于词袋的模型，以 LDA 模型[26]为代表。该类方法基于"文档-单词"的共现频率特征抽取主题，但忽视词汇之间的上下文语义关系。第二类是基于预训练词向量嵌入的模型，代表模型有 Top2Vec 等。该类方法假定主题相似的文档在语义空间中位置相近，但基于密度的文档聚类与基于中心的主题词采样产生矛盾，会导致误采样。BERTopic [27]模型采样基于文档集合的 C-TF-IDF 算法，从每个簇中选取词项来构建主题，克服了上述问题。

3 研究框架

本文研究的总体架构如图 3-1 所示。研究框架分为两大模块：数据获取模块，数据分析模块。

（1）数据获取模块的功能是数据采集和数据清洗。数据采集借助网站 API、搜索结果导出等方法，获取文献的链接列表，并进一步得到包含 DOI、标题、摘要等信息在内的元数据以及 altmetrics 信息；数据清洗主要是去除不完整的数据，并基于 DOI 对文献元数据与 altmetrics 数据进行匹配。最终获取来源 bioRxiv 与 medRxiv 的预印本数据、来源 Scopus 的期刊数据，并从 altmetric.com 网站获取每条文献数据对应的 altmetrics 信息。

（2）数据分析模块的功能是模型构建和热点对比。模型构建是对经过预处理的预印本与期刊数据分别构建 BERTopic 模型，调整参数，获取二者各自的主题分布情况；热点对比主要是通过二维象限主题对比法，从热度、影响两个维度，分析预印本、期刊的热点主题，并对比其相似与不同之处，为预印本与期刊的合作提出建议。

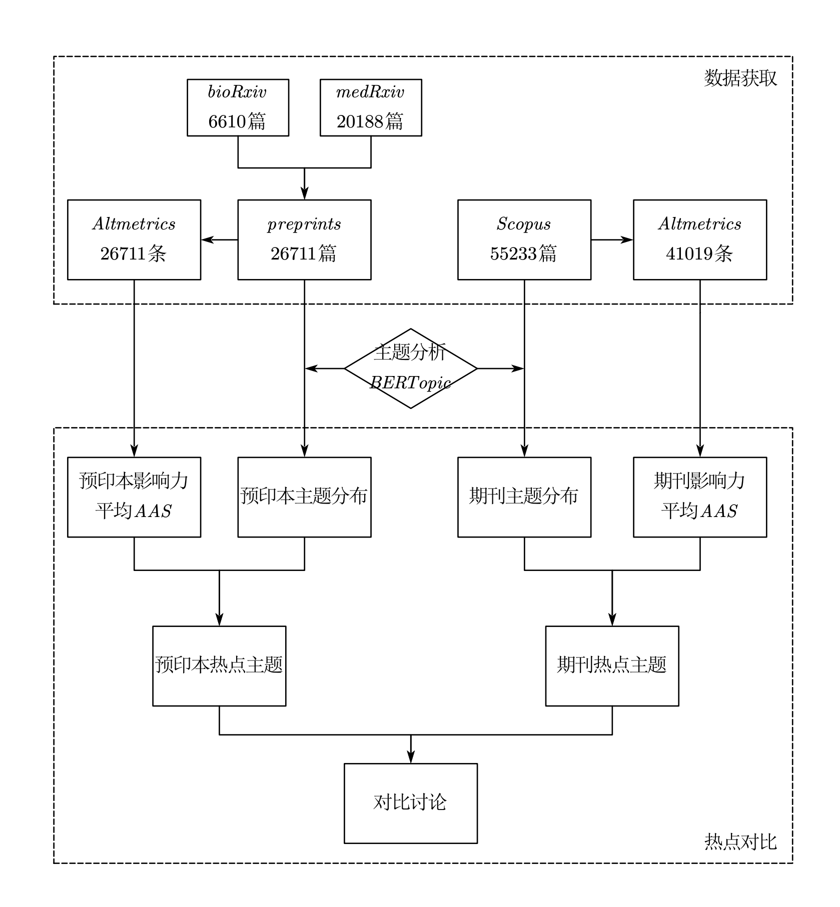


图 3-1 期刊预印本热点主题分析研究架构

4 研究方法

本文中的研究方法主要由两个部分组成。第一，在数据分析中的主题分析部分选择 BERTopic 模型。第二，在对期刊和预印本的热点主题比较分析中使用二维象限主题对比法。

4.1 BERTopic 模型

BERTopic 模型如图 4-1 所示。该模型弥补了基于词袋的主题模型忽视词汇上下文语义关系的缺点，又解决了基于预训练词嵌入模型使用不同方法进行文档聚类与主题采样导致的误采样问题，使得结果主题内的词项在语义上更加相关。该方法主要分为四个模块，分别是嵌入模块、降维模块、聚类模块、主题表征模块。

（1）嵌入模块主要目的是通过预训练语言模型来获得文档嵌入向量，这里使用默认的 BERT 预训练模型，即 Sentence-BERT 架构；

（2）降维模块用于降低嵌入的维度，方便后续聚类，这里采用的是一种非线性降维算法 UMAP ，可以保留更多局部特征，且速度较快；

（3）聚类模块采用基于密度的聚类算法 HDBSCAN，对降维后的文档向量进

行聚类，方法原文中把降维与聚类算作一个模块，此处分开介绍以便展示；

（4）模型的主题表征采用修改后的 TF-IDF 算法，即 C-TF-IDF，该算法基于上一步的聚类结果，为每个文档集群分配一个主题，如公式（4-1）所示[27]。

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t}) \quad (4\text{-}1)$$

其中词项的频率 $tf_{t,c}$ 指的是词项 t 在文档集群 c 中的频率，A 为每个集群的平均词数。通过 BERTopic 方法分别代入预印本与期刊的标题摘要文本，得到各自的文档-主题分布与主题-词项分布情况，以供后续对比分析。

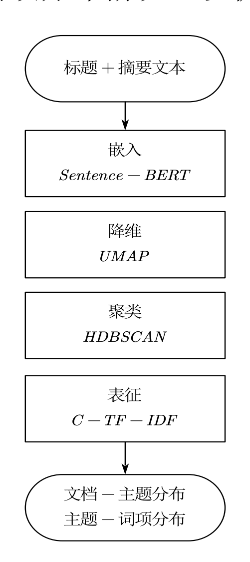


图 4-1 BERTopic 方法的主要模块

4.2 二维象限主题对比法

本文提出了二维象限主题对比法，可以直观地判断主题的类型，如图 3-3 所示。二维象限法曾由 Wang 等人用于期刊话语权研究[28]。本文基于具体情况提出了二维象限主题对比法，用于主题的对比。本文将主题的影响度和热度作为衡量指标，分别为四个象限的主题进行定义。第一象限为高影响高热度的“热点主题”，即发展较为成熟，有大量文献支持与影响力；第二象限为“新兴主题”，即相关研究大量增长，而新文献的影响力尚且不高；第三象限为“普通主题”，该象限的主题数量最多，但影响力与文献数量都一般；第四象限为“经典主题”，即文献数量不多，但影响力较大。当一个主题热度增加，大量研究出现，就会由“普通主题”跃迁到“新兴主题”；随着时间的推移，一些文献得到大量关注，主题的影响力提高，转变为“热点主题”；随着研究的变化，热点主题也会热度下降，一些仍有影响力的文献则精炼为“经典主题”，其他的则回归普通主题之中。

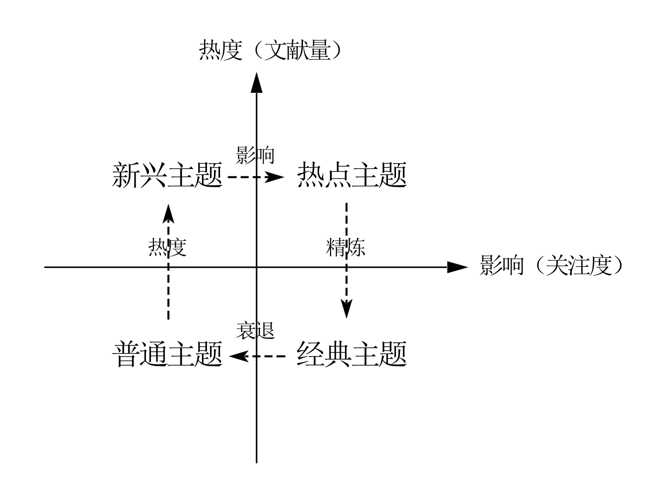


图 4-2 “热度-影响”二维象限法

为了将不同主题放入二维象限图中，本文提出公式（4-2）来对热度、影响两个维度的数值进行归一化。热度定义为该主题文献的数量，影响即该主题所有文献的平均关注度分数 AAS（Altmetric Attention Score），归一化采用的方法是“减去平均数，除以极大值”，使得数值分布在 [-1,1] 之间。

$$n = \frac{s-\bar{s}}{\max(s-\bar{s})} \quad (4\text{-}2)$$

5 实验结果

5.1 数据介绍

本文的研究对象为 2020 年至 2022 年 COVID-19 相关的期刊与预印本，期刊来源于 Scopus 数据库，预印本来源于 bioRxiv、medRxiv 数据库，altmetrics 信息获取自 altmetric.com 网站。采集时间为 2023 年 4 月 5 日。

预印本数据 bioRxiv 官网上 COVID-19 相关文献合集[1]，通过爬虫获取该合集中所有文献的网址、DOI 号，再借助官网的 API 获取所有文献的标题、摘要、作者等信息；期刊数据来源于 Scopus 数据库，以“COVID-19”为检索词在标题、摘要中检索，并通过筛选器进行精炼搜索，获取 COVID-19 相关的已发表期刊文章元数据；预印本与期刊对应的 altmetrics 信息获取自 altmetric.com 网站提供的 API，提供文献 DOI 号逐一匹配获取。去除无效信息后，最终获取预印本数据 26711 篇、期刊数据 55233 篇，以及各自对应的 altmetrics 信息。部分期刊信息齐全但无法查询到 altmetrics 信息，即未在社交媒体中提及，AAS 分数记为 0。

[1] 文献合集：https://connect.bioRxiv.org/relate/content/181

表 5-1 预印本与期刊文献元数据表

来源	数量	DOI	标题	摘要	Altmetrics
bioRxiv	6610	√	√	6604	√
medRxiv	20188	√	√	20107	√
Scopus	55233	√	√	√	41019

注：不齐全的地方填入具体数字，“ √ ”表示齐全

5.2 预印本热点主题分析

(1) 预印本高热度主题分析

将预印本数据代入模型自动聚类后，共获得 60 个主题。首先以每个主题的文献数量作为热度指标，列出预印本前 10 个高热度研究主题。如表 5-1 所示。

数量最多的主题 0 是“病毒传播模型”（Transmission-Model），根据相关词项与文档可以判断，该主题偏向宏观数据统计、传播模型建立与预测；主题 1 是“预防与检测”（Testing）相关，以核酸检测等为主；主题 2“基因组”（Genome）是对病理机制研究，该主题还包含大量病毒变异相关词项，可见病毒变异是该主题兴起的重要原因；主题 3 是“心理学”（Psychology）相关文献，主要涉及疫情期间的心理健康、精神压力等问题；主题 4 是“药物”（Drug）相关文献，主要涉及抗病毒药物的研制；主题 5 是“疫苗接种态度”（Vaccine-Attitude），涉及接种疫苗的犹豫、对疫苗的接受程度等研究；主题 6“蛋白质”（Protein）与主题 2 同样属于病毒机制研究，涉及蛋白质受体、刺突蛋白等；主题 7 是“疫情数据统计”（Statistics），即疫情期间的死亡率等数据统计与分析；主题 8 同样涉及病毒传播，但更偏向社会层面，命名为“病毒传播-社会”（Transmission-Social），讨论从传播途径控制病毒的传播，涉及口罩、空气传播等关键词；主题 9 同样为疫苗相关，但偏向疫苗研制，命名为“疫苗研制”（Vaccine-Development）。

表 5-1 预印本高热度主题

ID	主题	释义	词项	计数
0	Transmission-Model	病毒传播（偏统计、建模）	model, epidemic, number, cases, time, countries, transmission, measures, spread, population	2569

1	Testing	预防与检测（如核酸检测）	igg, antibodies, antibody, seroprevalence, infection, positive, igm, test, children, anti	1485
2	Genome	基因组相关病理机制（包含病毒变异）	mutations, genome, variants, sequences, sequencing, genomic, variant, mutation, lineages	1393
3	Psychology	心理学（疫情期间心理健康）	mental, health, mental health, anxiety, depression, social, psychological, care, stress, symptoms	991
4	Drug	药物等治疗手段	mpro, antiviral, drug, protease, inhibitors, drugs, activity, compounds, replication, treatment	591
5	Vaccine-Attitude	疫苗相关（偏接种态度）	vaccine, vaccination, hesitancy, vaccinated, vaccines, vaccine hesitancy, among, participants, uptake, acceptance	563
6	Protein	蛋白质相关病理机制	ace2, binding, protein, spike, rbd, receptor, spike protein, domain, affinity, receptor binding	553
7	Statistics	数据统计（死亡率等）	deaths, mortality, excess, countries, death, age, rates, fatality, population	453
8	Transmission-Social	病毒传播（社会层面）	mask, masks, air, aerosol, transmission, face, airborne, aerosols, droplets, ventilation	417
9	Vaccine-Development	疫苗相关（偏疫苗研制）	dose, vaccine, vaccination, antibody, bnt162b2, mrna, igg, anti, responses, second	356

参考 PubMed 网站对 COVID-19 文献的分类标准[2]，结合数据的具体情况，进一步归纳主题，以便分析规律。主题 0、主题 8 可归纳为“传播”（Transmission），即对病毒传播规律的研究；主题 1 的核酸检测，主题 5、主题 9 的疫苗都可以作为预防的一个部分，可归类为“预防”（Prevention）；主题 2 的基因组与主题 6 的蛋白质可归类为“机制”（Mechanism），即病毒原理研究；主题 3 的心理学与主题 7 的统计可归类为“社会”（Society），即社会层面的统计与其他研究；主题 4

[2] 分类标准：https://pubmed.ncbi.nlm.nih.gov/help/#covid19-articles

药物研制可视作治疗的一部分，可归类为“治疗”（Treatment）。

总体来看，预印本的高热度主题中，病毒的“传播”、“预防”、“机制”的研究偏多。“传播”中建模预测最多，社会层面防控其次；“预防”中核酸检测最多、疫苗其次；“机制”中基因组及病毒变异最多，蛋白质其次。“治疗”、“社会”层面研究偏少。

(2) 预印本高影响主题分析

以每个主题的平均 AAS 分数为影响指标，列出预印本前 10 个高影响主题，如表 5-2 所示。这些主题关注度与讨论度较高，可以体现研究的重要性。为了方便区分，保留实验结果中的主题号作为 ID。可以发现前 10 个高影响主题与高热度主题重合度较小，但存在部分内容相似的主题。

主题 39、主题 9、主题 56 均为疫苗相关主题，根据对相关词项与对应文档的观察，主题 9 更偏向二次接种（Vaccine-Second），主题 56 更偏向病毒变体（Vaccine-Variant）；主题 27、主题 10 均为病毒变异相关研究，其中主题 27 面向疫苗研制（Variant-Vaccination），主题 10 更综合一些；主题 26 为“临床治疗”（Clinical），相关词包含临床治疗的药物与有机化合物；主题 41 为“免疫”（Immune），涉及记忆细胞、抗体等免疫相关病理机制；主题 54 为“动物载体”（Transmission-Animal），即病毒传播的动物载体；主题 55 偏向“神经学科”（Neurology），涉及大脑、神经研究，与治疗、心理学均有关联；主题 32 为“心理学”（Psychology）相关，偏向认知心理学，即社会层面的心理健康研究。

表 5-2 预印本高影响主题

ID	主题	释义	词项	分数
39	Vaccine-Development	疫苗相关（mRNA 等疫苗研发）	dose, bnt162b2, vaccine, vaccination, effectiveness, infection, ci, vaccinated, mrna	948.31
27	Variant-Vaccination	病毒变异（面向疫苗研制）	omicron, delta, variant, ba, infection, vaccination, omicron variant, ci, compared	551.22
26	Clinical	治疗（临床、药物治疗）	hcq, hydroxychloroquine, patients, treatment, trials, chloroquine, group, clinical, cq, azithromycin	419.12
10	Variant	病毒变异（综合）	omicron, ba, variant, variants, neutralizing, neutralization, delta,	334.3

			antibody, omicron variant	
9	Vaccine-Second	疫苗相关（二次接种）	dose, vaccine, vaccination, antibody, bnt162b2, mrna, igg, anti, responses, second	325.34
41	Immune	病理机制（记忆细胞、抗体等免疫相关）	memory, memory cells, cells, cell, specific, antibodies, responses, vaccination, immune, antibody	321.64
56	Vaccine-Variant	疫苗相关（病毒变体）	vaccine, variants, vaccination, neutralizing, antibody, infection, protection, mrna, vaccines, antibodies	315.42
54	Transmission - Animal	动物（病毒传播的载体）	cats, deer, animals, animal, humans, species, mink, dogs, transmission, tailed deer	290.56
55	Neurology	神经学科	brain, neurological, neurons, microglia, infection, cns, cells, neuroinflammation, neurological symptoms, nervous	240.36
32	Psychology	心理学（偏心理与认知）	neurological, cognitive, patients, symptoms, psychiatric, brain, disorders, long, acute, mental	203.1

与前一小节类似，对预印本高影响主题进行归类。主题 39、主题 9、主题 56 的疫苗相关研究归类为“预防”（Prevention）；主题 27、主题 10 的病毒变异相关与主题 41 的免疫相关，均可归类为“机制”（Mechanism）；主题 26 的临床治疗、主题 55 的神经科学都可以计入“治疗”（Treatment）；主题 54 归类为“传播”（Transmission）；主题 32 的心理学方面可归类为“社会”（Society）。

总体来看，预印本的高影响力主题主要涉及“预防”、“机制”两个大类，其中“预防”主要是关于疫苗的研究，而“机制”主要是病毒变异相关。对比高热度主题，病毒“传播”相关研究虽然数量较多，但文献的平均影响力不高。“治疗”层面的研究中，药物研制主题热度高，临床治疗主题影响大。“社会”层面的研究在热度与影响方面都不突出，主要是作为其他研究的拓展与补充而存在。

5.3 期刊热点主题分析

(1) 期刊高热度主题

将期刊数据代入模型自动聚类后，共获得 122 个主题。首先以每个主题的文献数量作为热度指标，列出期刊前 10 个高热度研究主题。可以发现热点主题与

预印本差别较大。最多的主题为“在线教育”（Online Learning），即社会层面的线上教育方面的研究与介绍；主题 1 与预印本热点主题一致，是“病毒传播”（Transmission），可见关于病毒传播过程的统计与研究普遍较多；主题 2 与主题 3 均与疫苗有关。主题 2 偏向疫苗的原理与研制，命名为“疫苗研制”（Vaccine-Development）；主题 3 偏向民众对疫苗的态度，命名为“疫苗接种态度”（Vaccine-Attitude）；主题 4 为“蛋白质”（Protein），即病毒蛋白质相关机制研究；主题 5 为“在线健康”（Telehealth），与在线教育一样兴起于对线下活动的管控，包含线上问诊等相关统计与研究；主题 6 为“运动”（Exercise），与在线教育、在线健康一样，属于社会层面研究，即面向民众疫情期间的生活质量；主题 7“检测”（Testing）是以核酸检测为主的预防手段相关研究；主题 8“社会”（Society）则是对疫情期间社会层面的经济、政治方面的总体研究；主题 9“污染”（Pollution）主要是空气污染，与病毒传播的研究相关。

表 5-3 期刊高热度主题

ID	主题	释义	词项	计数
0	Online Learning	在线教育	students, online, education, teaching, teachers, student, training, virtual, online learning	1994
1	Transmission	病毒传播	epidemic, cases, mobility, number, spread, transmission, countries, time, models	1711
2	Vaccine-Development	疫苗（疫苗研制）	antibody, antibodies, vaccination, vaccine, dose, anti, bnt162b2, neutralizing, infection	1467
3	Vaccine-Attitude	疫苗（接种态度）	vaccination, hesitancy, vaccine hesitancy, vaccines, vaccinated, acceptance, uptake, among, trust	1177
4	Protein	蛋白质	protein, compounds, protease, molecular, mpro, docking, drug, ace2, inhibitors	1107
5	Telehealth	在线健康	telehealth, care, visits, patient, services, virtual, person, video, patients	1003
6	Exercise	运动健康	physical, food, activity, pa, eating, weight, exercise, lockdown, changes	998
7	Testing	预防与检测	hcws, infection, transmission, workers, risk, testing, respiratory, positive, staff	728
8	Society	社会（政	health, crisis, political, social, policy,	714

		策、经济）	public, economic, world	
9	Pollution	污染（空气污染）	pm2, pollution, no2, air quality, lockdown, air pollution, emissions, concentrations, pollutants	699

期刊的高热度主题与预印本的区别较大，可以发现主题 0、主题 5、主题 6、主题 8 都可以归类到“社会”（Society），显然相较预印本只有统计、心理等层面的研究要丰富许多；主题 1、主题 9 可以归类到“传播”（Transmission），涉及病毒的传播与空气质量的管控；主题 2、主题 3 的疫苗相关研究、主题 7 核酸检测相关，都可归类到“预防”（Prevention），疫苗相关研究的数量尤其多；主题 4 蛋白质相关可以归入“机制”（Mechanism），相较预印本的包含基因组、蛋白质、免疫系统等在内的机制研究，期刊的病理机制研究在数量上占比较少。总而言之，在热度方面，期刊的主题更偏向社会层面，且分类更丰富。

(2) 期刊高影响主题

以每个主题的平均 AAS 分数为影响指标，列出期刊前 10 个高影响主题。与预印本高影响主题类似，主题 107、主题 2 都是影响力较高的疫苗相关主题，其中主题 107 更偏向疫苗研制中的试验（Vaccine-Trial），主题 2 比较综合，可命名为“疫苗研制”（Vaccine-Development）；主题 73 则涉及疫情期间的“补给”（Supplement），与治疗层面有关，不过更偏向社会层面的研究；主题 27 为“临床治疗”（Clinical），相关词项包含了许多临床药物与有机物名称；主题 48、主题 82、主题 57 均为病毒传播相关主题，其中主题 48 更强调口罩等社会层面的管控（Social），其他则偏向空气污染（Contamination）、气溶胶（Aerosol）等传播渠道研究；主题 22 为“蛋白质”（Protein）相关病理机制研究，涉及了基于病毒机制进行的疫苗研制；主题 28 为“神经科学”（Neurology）相关，与预印本高影响主题中的神经科学一致。

表 5-4 期刊高影响主题

ID	主题	释义	词项	分数
107	Vaccine-Trial	疫苗（研制试验）	vaccination, vaccines, dose, vaccinated, safety, trial, placebo, efficacy, group	278.28
73	Supplement	补给（缺乏）	deficiency, vitamin deficiency, zinc, supplementation, levels, serum	274.33

27	Clinical	治疗（临床）	hydroxychloroquine, treatment, qtc, chloroquine, patients, clinical, azithromycin, trials, prolongation	245.93
48	Transmission-Social	传播（口罩等）	masks, wearing, face, face masks, mask wearing, face mask, faces, wear, public	243.18
82	Transmission-Contamination	传播（空气污染）	air, samples, contamination, environmental, rna, transmission, surface, airborne, aerosols	215.62
2	Vaccine-Development	疫苗（接种研制）	antibody, antibodies, vaccination, vaccine, dose, anti, bnt162b2, neutralizing, infection	190.02
57	Transmission-Aerosol	传播（气溶胶）	ventilation, airborne, indoor, droplets, transmission, aerosol, droplet, aerosols, risk	173.09
22	Protein	蛋白质（病毒变异、疫苗研制）	neutralizing, antibodies, spike, binding, antibody, variants, protein, vaccine, vaccines	153.03
28	Neurology	神经科学	manifestations, brain, neurological manifestations, patients, acute, encephalopathy, symptoms, neurologic, mri	131.43

主题 107、主题 2 的疫苗相关研究可以归类为“预防”（Prevention）；主题 73 的补给可以归入“社会”（Society）；主题 27 的临床治疗、主题 28 的神经科学可以归入“治疗”（Treatment）；主题 48、主题 82、主题 58 均可归入“传播”（Transmission）；主题 22 的蛋白质可以归入“机制”（Mechanism）。

可以发现，“传播”层面的研究，在预印本高影响主题中不够显著，却在期刊中占据很大比重；以疫苗为主的“预防”，在期刊、预印本中都属于高影响的主题；“治疗”层面预印本与期刊一致，都是临床相关主题影响力较高；蛋白质等病毒的“机制”相关研究在在期刊中影响力较小；“社会”层面的研究在期刊中虽数量较多，但平均影响力不够高。

5.4 预印本与期刊热度-影响二维象限分析

为了更直观地展示不同主题在“热度-影响”两个维度的分布，这里对两个指标进行归一化处理后展示在二维坐标系上。

预印本主题如图 5-1 所示。可以发现大部分主题属于低影响低热度的“普通主题”，热度最高的“新兴主题”是“病毒传播模型”（Transmission-Model），该类研究在疫情期间大量增加，大多旨在预测病毒传播与疫情发展规律，但影响力不够高；影响最大的“经典主题”为“疫苗研制”（Vaccine-Development），疫苗相关研究包含许多主题，该细分主题影响力偏高一些。

高影响高热度的“热点主题”，有主题 10“病毒变异”（Variant）、主题 9“疫苗-二次接种”（Vaccine-Second）、主题 8“病毒传播-社会”（Transmission-Social）、主题 5“疫苗接种态度”（Vaccine-Attitude）。距离原点最远的为“病毒变异”（Variant），该主题在预印本中尤为突出，属于特有的热点主题；其次为“疫苗-二次接种”（Vaccine-Second），这两个主题热度与影响都较高。

期刊主题如图 5-2 所示。相较预印本主题分布，期刊的主题分布更加分散。高影响低热度的“经典主题”以“疫苗试验”（Vaccine-Trial）为代表，与预印本一样属于疫苗与“预防”主题；高热度低影响的“新兴主题”以“在线教育”（Online Learning）为代表，相较于预印本更偏向“社会”层面。高影响高热度的“热点主题”相较预印本数量更多，且存在明显远离原点的主题“疫苗研制”（Vaccine-Development），可见疫苗相关研究无论在预印本还是期刊中都是高热度高影响的“热点主题”。

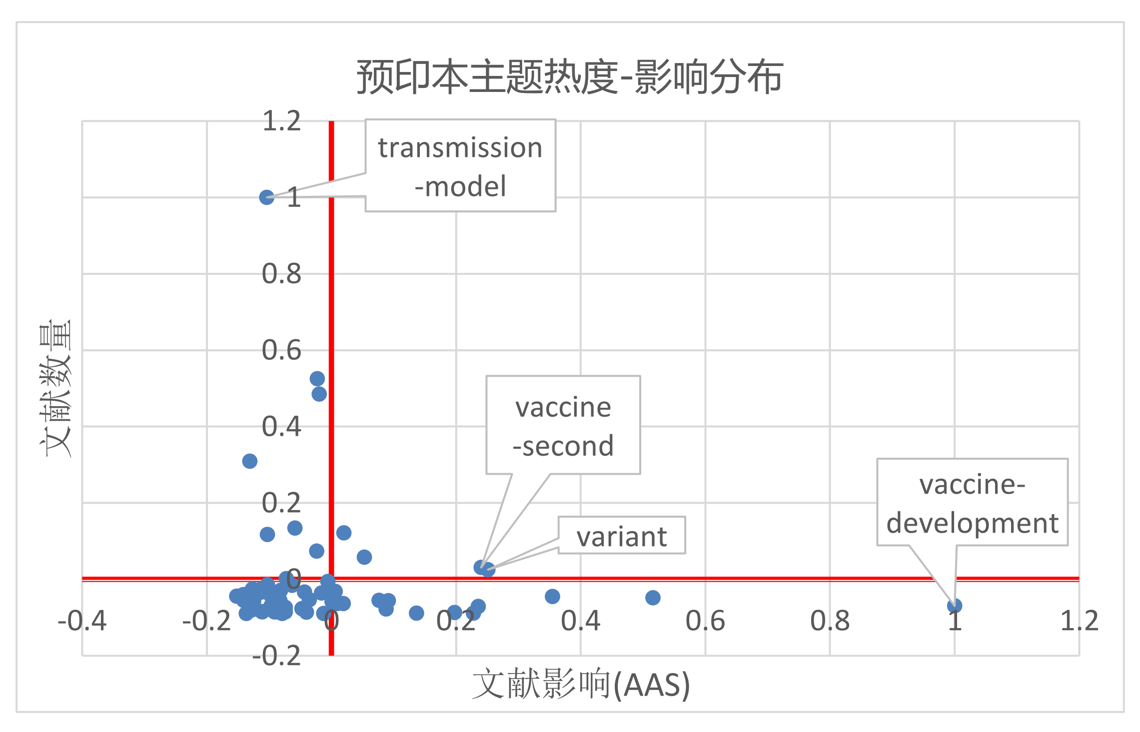


图 5-1 预印本主题热度-影响分布

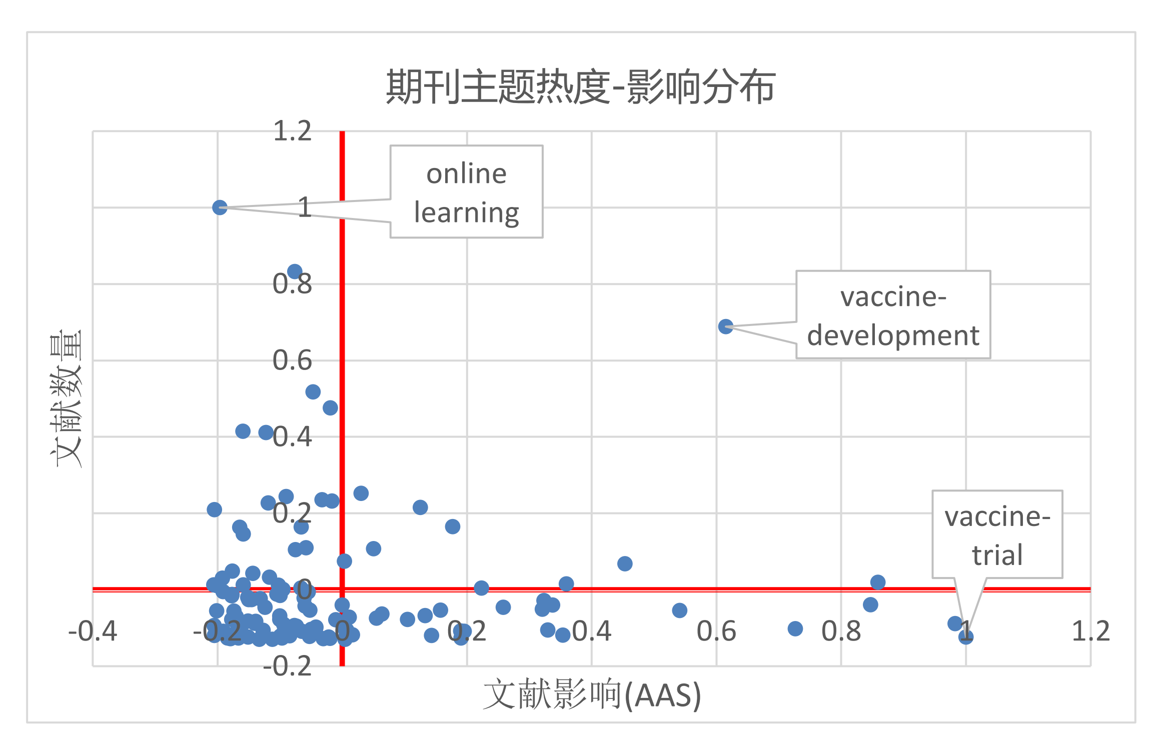


图 5-2 期刊主题热度-影响分布

实验中参考 PubMed 网站的分类标准，把主题分布结果归结为五大类，即“传播”（Transmission）、“预防”（Prevention）、“机制”（Mechanism）、“治疗”（Treatment）、“社会”（Society）五大类。五大主题并非严格区分，而是存在一定的关联，如研究“传播”规律对于“预防”有帮助；研究病毒“机制”有助于疫苗和药物的研制，分别对应“预防”与“治疗”；而民众对待“预防”的态度、心理健康方面的“治疗”、“传播”结果的统计等，又产生“社会”层面的研究主题。如图 5-3 所示。

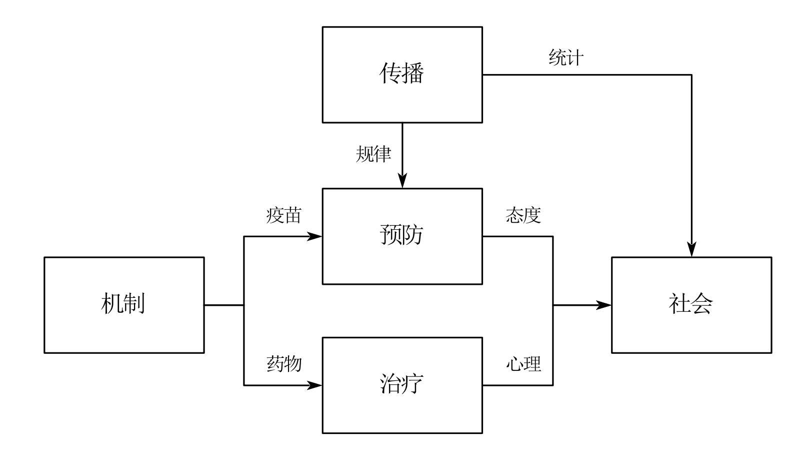


图 5-3 COVID-19 大类主题相互关系

6 总结讨论

本文采用 BERTopic 模型，以新冠疫情主题为例，分别对来自 bioRxiv、medRxiv 的预印本文献，来自 Scopus 的期刊文献，从热度、影响两个维度进行主题分析。研究获得预印本与期刊各自偏好的热点主题，并参考 PubMed 网站的分类标准进行归类，在对比中发现其潜在的规律，验证了预印本与期刊在研究内容方面的互补。通过归结出期刊与预印本在不同维度的热点主题，对比其联系与区别，可以得出以下结论。

（1）从热度上看，预印本主题更偏向病毒传播、预防、病理机制等医学层面，且“机制”（Mechanism）大类相较期刊的分类更细致，包含蛋白质、基因组等主题；期刊的主题兼具医学与社会层面，且“社会”（Society）大类的主题分类更细致，包含在线教育、在线健康等主题；

（2）从影响上看，以疫苗主题为代表的“预防”（Prevention）大类在预印本与期刊中都占据重要地位，尤其是在预印本文献中；“传播”（Transmission）相关主题则在期刊中影响力较大；

（3）预印本与期刊在热度和影响力方面有部分重回主题。例如疫苗相关研究，无论在热度维度还是影响维度、在预印本中还是期刊中，都是重要的研究热点主题，总体来看在影响力方面更加突出；

（4）预印本中有期刊中所没有的特有的热点主题。例如病毒变异相关研究在预印本主题中属于高影响高热度的热点主题，而在期刊中不够显著，可以视作预印本主题的特点所在。

总体来看，就新冠疫情主题而言，尽管期刊的主题较为全面，预印本仍有其特点与偏向。预印本主题更偏向理论与学术，而期刊主题更偏向实践与社会；预印本主题在原理方面分支更细致，而期刊主题在社会层面涉及面更广，二者在内容方面可以形成有效互补。对此，预印本平台可以更多地鼓励理论创新，扬长避短，发挥预印本速度快、形式灵活的优势，在内容方面与期刊进一步差异化以实现学术交流上的分工合作、优势互补。研究人员可以发挥预印本的长处，通过其了解最新的理论创新与思想火花，更好地发挥预印本在学术交流中的作用。

参考文献

[1] 陈雪飞, 张智雄, 黄金霞. 国际学术出版机构预印本政策分析 [J]. 数字图书馆论坛, 2017, (10): 8-14.

[2] Ross J S, Krumholz H M. Ushering in a New Era of Open Science Through Data Sharing The Wall Must Come Down [J]. Jama-Journal of the American Medical Association, 2013, 309(13): 1355-1356.

[3] 张智雄, 黄金霞, 王颖, et al. 国际预印本平台的主要发展态势研究 [J]. 数字图书馆论坛, 2017, (10): 2-7.

[4] 唐耕砚. 重构与再造：预印本平台对科学交流体系的影响 [J]. 科学学研究, 2021, 39(10): 1729-1735+1831.

[5] 唐耕砚, 蔡豪. 预印本平台的舆论治理困境与应对策略——基于“新型冠状病毒肺炎”事件的反思 [J]. 科学学研究, 2021, 39(04): 587-593.

[6] Perelman G. The entropy formula for the Ricci flow and its geometric applications [EB/OL]. 2002. https://arxiv.org/abs/math/0211159.

[7] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. 2019. https://arxiv.org/abs/1810.04805.

[8] 刘菊红. 自存档文章引用优势案例分析研究 [J]. 图书情报工作, 2008, 52(11).

[9] 徐诺, 苗秀芝, 程建霞. 预印本“大繁荣”对科技期刊编辑的启示 [J]. 编辑学报, 2019, 31(3): 282-285,289.

[10] 周阳. 国内外预印本系统调研与启示 [J]. 图书馆界, 2021, (03): 60-68.

[11] 汪庆, 任慧玲. 开放融合环境下预印本发展态势分析 [J]. 数字图书馆论坛, 2021, (08): 59-64.

[12] 刘静羽, 张智雄, 黄金霞, et al. 预印本服务中的质量控制方法研究 [J]. 数字图书馆论坛, 2017, (10): 15-19.

[13] 宋永辉, 马廷灿, 刘静羽. 志愿者参与下国际预印本平台学术质量控制方法调研与启示——以 arXiv 与 RePEc 平台为例 [J]. 中国科技期刊研究, 2023, 34(02): 119-126.

[14] Nicholson D N, Rubinetti V, Hu D B, et al. Examining linguistic shifts between preprints and publications [J]. Plos Biology, 2022, 20(2).

[15] Clyne B, Walsh K A, O'Murchu E, et al. Using preprints in evidence synthesis: Commentary on experience during the COVID-19 pandemic [J]. Journal of Clinical Epidemiology, 2021, 138: 203-210.

[16] Shi X, Ross J S, Amancharla N, et al. Assessment of Concordance and Discordance Among Clinical Studies Posted as Preprints and Subsequently Published in High-Impact Journals [J]. Jama Network Open, 2021, 4(3): e212110-e212110.

[17] 刘春丽, 盛南洪. bioRxiv 自存档的期刊论文多维度影响力优势实证研究 [J]. 信息资源管理学报, 2022, 12(04): 33-45.

[18] Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders [J]. Scientometrics, 2020, 124(3): 2687-2701.

[19] 匡登辉, 王丽婷. COVID-19 预印本的影响力分析及其发展启示 [J]. 数字图书馆论坛, 2020, (12): 45-50.

[20] 李爱花, 任慧玲, 张玢. 基于文献计量的新型冠状病毒肺炎研究现状分析 [J]. 数字图书馆论坛, 2020, (03): 2-8.

[21] Santos B S, Silva I, Lima L, et al. Discovering temporal scientometric knowledge in COVID-19 scholarly production [J]. Scientometrics, 2022, 127(3): 1609-1642.
[22] Jabalameli S, Xu Y, Shetty S. Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination [J]. International Journal of Disaster Risk Reduction, 2022, 80: 103204.
[23] 钟伟金, 李佳. 共词分析法研究(一)——共词分析的过程与方式 [J]. 情报杂志, 2008, (05): 70-72.
[24] 王志涛, 於志文, 郭斌, et al. 基于词典和规则集的中文微博情感分析 [J]. 计算机工程与应用, 2015, 51(8): 218-225.
[25] 吴睿. 面向微博文本的热词分析技术研究 [D]. 昆明: 昆明理工大学, 2019.
[26] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[27] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure [J]. 2022.
[28] Wang X. Evaluation of the discourse power in Chinese academic journals: A multi-fusion perspective [J]. Data and Information Management, 2023.

作者简介

杜潇霖，男，硕士，研究方向：情报学，E-mail：duxiaolin320325@163.com。
虞为，女，博士，副教授，研究方向：情报学，E-mail：yuwei@nju.edu.cn
陈俊鹏，男，博士，副教授，E-mail：Luckjp@163.com

基金项目

本文系国家社科基金项目“协同共治视角下图书馆推进开放科学的服务模式研究”（项目编号：21BTQ030）研究成果之一